MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2019/2020

TDS2101 – INTRODUCTION TO DATA SCIENCE

(All sections / Groups)

22nd OCTOBER 2019
9.00 a.m – 11.00 a.m

(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 6 pages with 4 Questions only excluding the cover page

2. Attempt **ALL** questions. All questions carry equal marks and the distribution of the marks for each question is given.

3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

(a) Describe the importance of **EACH** of the following *business* skills in a data scientist.

(i) Effective Communication [2 marks]
(ii) Industry Knowledge [2 marks]
(iii) Analytic Problem-Solving [2 marks]

(b) One of the stages in data science process is to *model the data*. What are the **THREE** key activities that are conducted at this stage ? [3 marks]

(c) State **ONE** challenge that might be faced by data scientists when dealing with Big Data. [1 mark]

QUESTION 2

(a) Table 1 represents the personal data of female employees in organization ABC. The personal details that are captured include name, monthly income, total number of cars that are registered under their name and whether their age is above 40 years old.

Table 1: Personal data of female employees in organization ABC.

Name	Monthly-Income	Number-of-Cars	Above-40
Alice	8000	3	Yes
Lydia	15000	5	Yes
Cindy	3000	1	No
Alicia	5000	2	No
Lavenia	4600	2	Yes
Melissa	7500	3	No

CONTINUED...

(i) Apply min-max normalization to normalize the values of **Monthly-Income** and **Number-of-Cars** so that the values fall within the range of [0,1]. Draw a similar template as below in your Answer Booklet. The new columns namely *Norm-Income* and *Norm-Cars* will store your workings and answers. Round your answers to two decimal places.

TEMPLATE

Monthly-Income	**Norm-Income**
8000	
15000	
3000	
5000	
4600	
7500	

Number-of-Cars	**Norm-Cars**
3	
5	
1	
2	
2	
3	

[3 marks]

(ii) Why is it essential to normalize the data before applying data mining algorithm to the data? [1 mark]

(b) Differentiate between descriptive question and exploratory question. Provide **ONE** example of descriptive question and **ONE** example of exploratory question. [4 marks]

(c) What are the **TWO** libraries that should be installed to ensure that Python is usable for data analysis? [2 marks]

CONTINUED...

QUESTION 3

(a) Table 2 records the symptoms that are observed by a doctor when he is consulting patients in a clinic. The values "1" and "0" indicate the presence and absence of the symptoms respectively. A total of three patients have visited the clinic on Sunday morning.

Table 2: Doctor's observations on three patients

	Flu	Sore Throat	Swollen Eyes	Swollen Feet	Rashes	Cough
Patient-A	0	1	0	1	1	0
Patient-B	1	1	1	0	0	1
Patient-C	0	0	0	1	1	0

(i) State the similarity measure that can be applied to identify the similarity of the patients that have visited the clinic on Sunday morning, given the symptoms. [1 mark]

(ii) Based on your answer in Question 3(a)(i), apply the similarity measure to identify the patients that have similar symptoms. Show your workings and round your answers to two decimal places. [4 marks]

(b) Document-based store is a type of NoSQL database. Show an example of how the data is stored in document-based store. [2 marks]

(c) Figure 1 shows a confusion matrix that is generated from applying algorithm X to a data in order to predict whether an individual will contract flu.

		Predicted	
		Flu: Yes	Flu: No
Actual	Flu:Yes	60	40
	Flu: No	30	50

Figure 1: Confusion matrix

(i) State the formulas for precision and recall. [2 marks]

(ii) Calculate the precision and recall for algorithm X. Round your answers to two decimal places. [1 mark]

CONTINUED...

QUESTION 4

Assume that R's working directory is set to the right location to retrieve the file and relevant packages have been loaded into current R working environment.

(a) Figure 2 shows the records of children in a pre-school. These records are stored in a comma separated value file named *child.csv*.

```
Name, Age, Height, Weight
Ali, 5, 110.1, 18.6
Mimi, 6, 100.7, 15.2
Fiona, 5, 105.0, 17.8
Alan, NA, 95.8, 15.5
Suzi, NA, 85.3, 14
Bernard, 6, 121.6, 20.4
Alix, 5, 102.9, 17.9
Ahmad,6, 120.3, 19.5
Johnny, NA, 100.3, 15.3
Henry,5, 105.9, 17.9
```

Figure 2: child.csv

Using R, create a function with no argument list and assign it to variable *one*. This function loads the *child.csv* to R working environment and performs the following three operations.

- It counts the number of rows having "NA" as one of its values.
- It replaces "NA" with the value 4.
- It displays the content of *child.csv* after the replacement operation.

The desired output is given as in Figure 3.

```
> one()
[1] "There are  3  rows having NA as one of its values"
       Name Age Height Weight
1       Ali   5  110.1   18.6
2      Mimi   6  100.7   15.2
3     Fiona   5  105.0   17.8
4      Alan   4   95.8   15.5
5      Suzi   4   85.3   14.0
6   Bernard   6  121.6   20.4
7      Alix   5  102.9   17.9
8     Ahmad   6  120.3   19.5
9    Johnny   4  100.3   15.3
10    Henry   5  105.9   17.9
```

Figure 3: The desired output for Question 4(a)

[3 marks]

CONTINUED...

(b) Figure 4 shows the updated records of children in a pre-school. These records are stored in a comma separated value file named *child-update.csv*.

```
Name, Age, Height, Weight
Ali,5,110.1,18.6
Mimi,6,100.7,15.2
Fiona,5,105,17.8
Alan,4,95.8,15.5
Suzi,4,85.3,14
Bernard,6,121.6,20.4
Alix,5,102.9,17.9
Ahmad,6,120.3,19.5
Johnny,4,100.3,15.3
Henry,5,105.9,17.9
```

Figure 4: child-update.csv

Using R, create a function with no argument list and assign it to variable *two*. This function loads the *child-update.csv* to R working environment, then calculates the mean height and mean weight of the children with respect to their age. The desired output is given as in Figure 5.

```
> two()
  Age   Mean-Height   Mean-Weight
1   4        93.800      14.93333
2   5       105.975      18.05000
3   6       114.200      18.36667
```

Figure 5: The desired output for Question 4(b)

[5 marks]

CONTINUED...

(c) Figure 6 shows the profits (in RM) earned by a salesman. The profits are recorded according to years and quarters. For example, *2013* refers to year 2013 while *Q1-2018* refers to the first quarter of the year 2018.

```
Year, Profit
2013, 1000
2014, 5000
2015, 2500
2016, 6000
2017, 4800
Q1-2018, 520
Q2-2018, 620
Q3-2018, 150
Q4-2018, 100
```

Figure 6: profit.csv

Using R, write the code to build a line graph that shows the **yearly** profit earned by the salesman from year 2013 to year 2018. The desired output is given as in Figure 7.

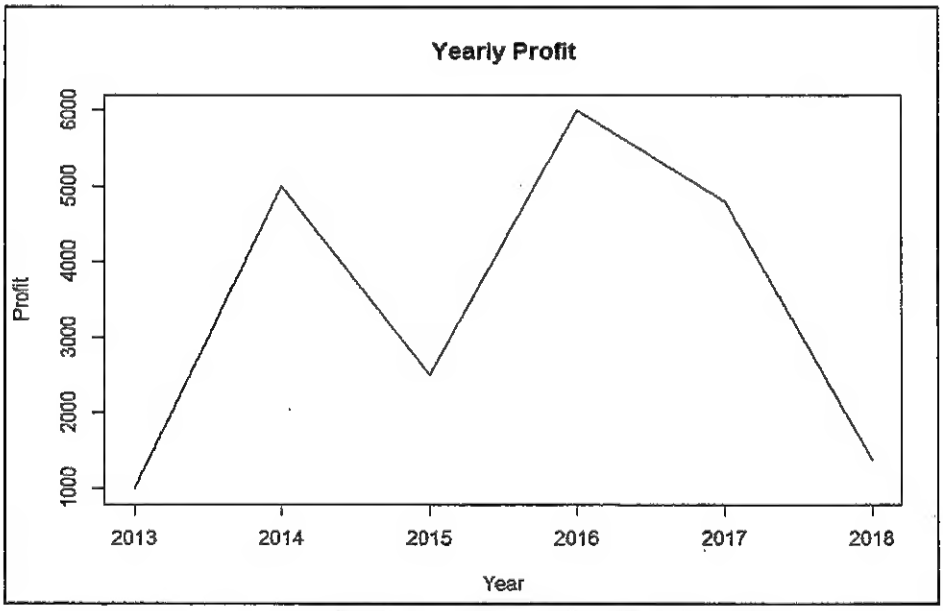

Figure 7: The desired output for Question 4(c)

[2 marks]

END OF PAGE.